DSS

Data & Storage Services

CERN IT Department

Big Data Management

Dirk Duellmann
CERN IT

Workshop on HPC and Super-computing for Future Science Applications

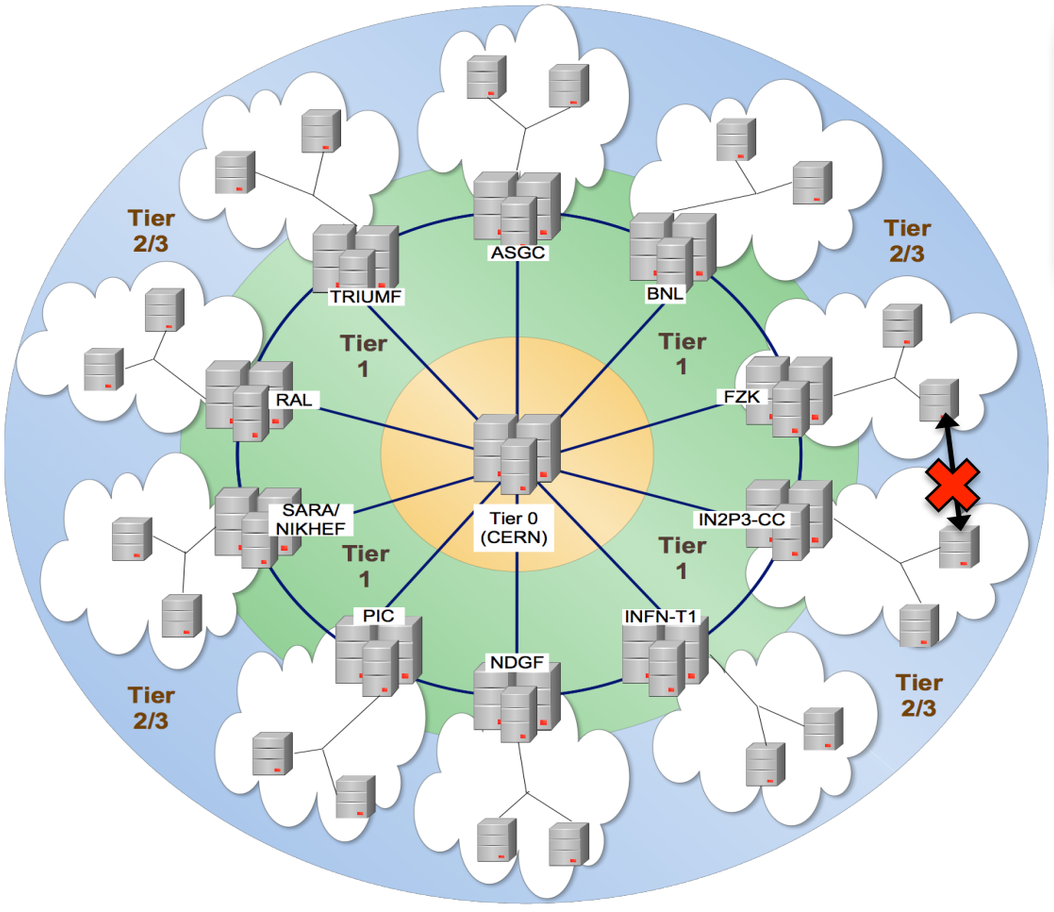
Brookhaven National Laboratory, June 2013

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Slide: Borut Kersevan

WLCG tiered structure

- The LHC experiments rely on **distributed computing resources:**
  - **WLCG - a global solution, based on the Grid technologies/middleware.**
    - distributing the data for processing, user access, local analysis facilities etc.
    - at time of inception envisaged as the seed for global adoption of the technologies.
  - **Tiered structure:**
    - Tier-0 at CERN: the central facility for data processing and archival,
    - 11 Tier-1s: big computing centers with high quality of service used for most complex/intensive processing operations and archival,
    - ~140 Tier-2s: computing centers across the world used primarily for data analysis and simulation.
  - **WLCG and LHC computing a big success in Run 1!**
    - **Computing was not a limiting factor for the Physics program of the LHC experiments.**
    - **Many thanks to our Grid sites for their excellent performance and contributions!**

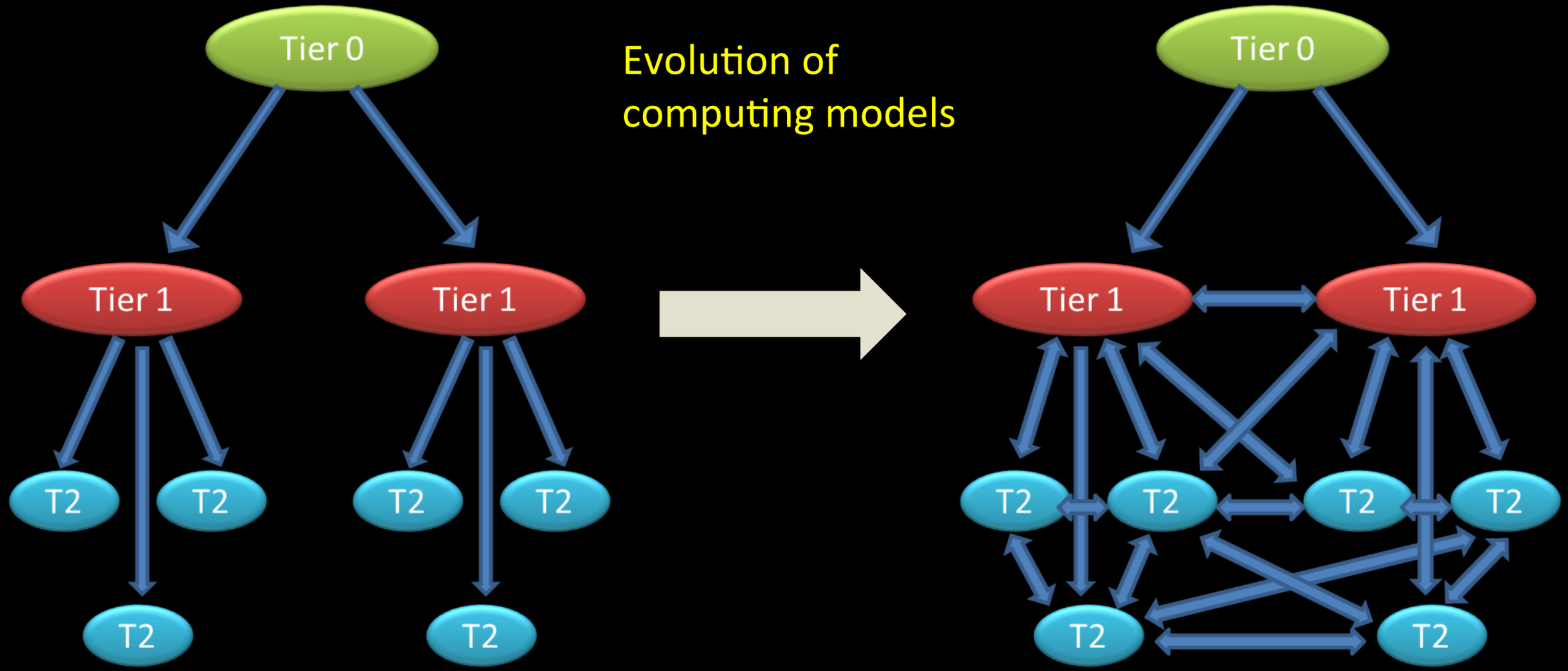


Capacity:

- ~350,000 CPU cores
- ~200 PB of disk space
- ~200 PB of tape space

- **Hierarchical tier organization** based on **MONARC** (**MO**DELS **OF** **N**ETWORKED **A**NALYSIS **AT** **R**EGIONAL **C**ENTERS) network topology
- In **ATLAS** sites are grouped into **clouds** for organizational reasons
- Possible communications:
  - Optical Private Network
    - T0-T1
    - T1-T1
  - National networks
    - Intra-cloud T1-T2
- Restricted communications: General public network
  - Inter-cloud T1-T2
  - Inter-cloud T2-T2

Computing model evolution



Hierarchy

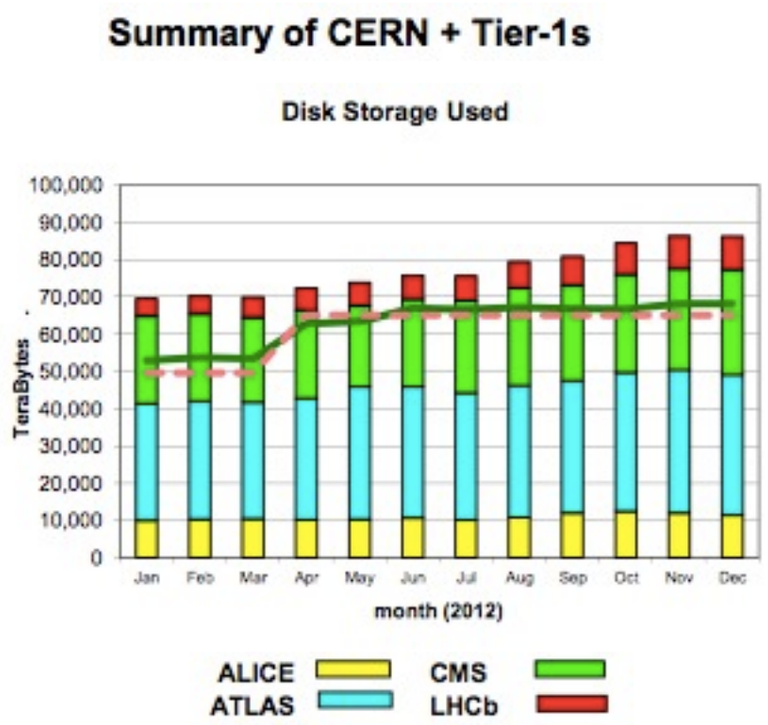
Mesh

WLCG Worldwide LHC Computing Grid

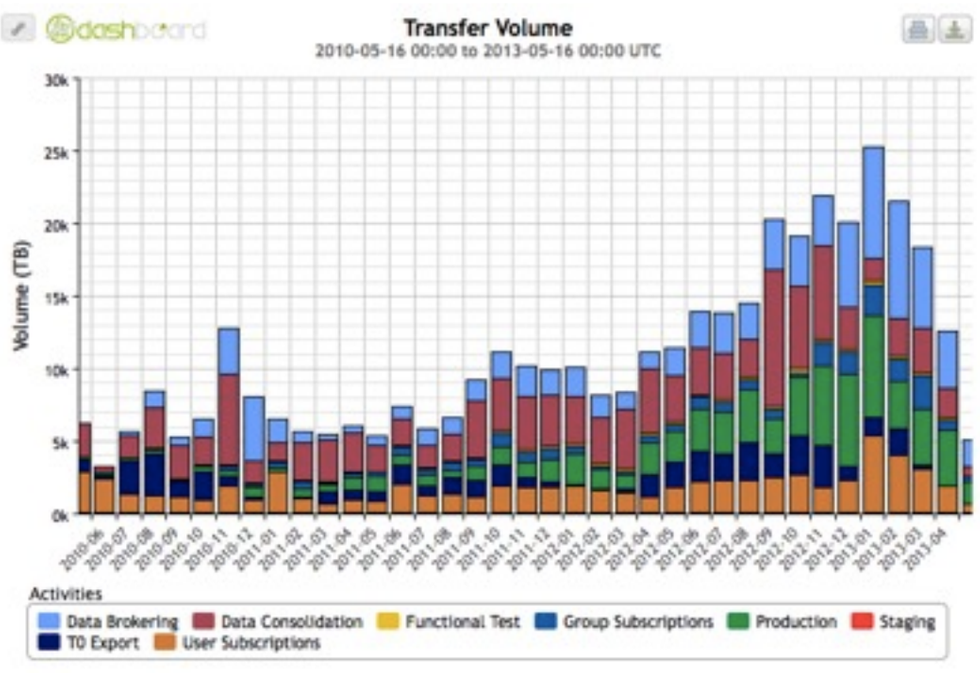
Processing the Experiment Big Data

Slide: Borut Kersevan

- The simplest solution in processing our data is **using the data affinity** for the jobs:
  - data is staged to the site where the compute resources are located and data access by analysis code is from local, site-resident, storage.
- **However:**
  - In our distributed computing environment we do not have enough disk space to host all our data on every WLCG site:
    - Thus we distribute (pre-place) our data across our sites.
  - The popularity of certain data sets is very hard to judge in advance:
    - Thus the computing capacity at a site might not match the demand for certain data sets.
- **Different approaches are being implemented:**
  - **Dynamic or on-demand data replication:**
    - **Dynamic:** If certain data is popular on one site (i.e. processed often), make additional copies on sites with spare CPU (and disk) capacity. (e.g. ATLAS PD2P service).
    - **On-demand:** The popular data on one site can be copied locally by jobs in another site (XRootD, HTTP federations).
  - **Remote access:**
    - The popular data on one site can be remotely accessed from jobs in another site (XRootD, HTTP federations).
  - **Both approaches have the underlying scenario that puts the WAN between the data and the executing analysis code.**
    - Inserting the WAN is a change that potentially requires special measures to ensure the smooth flow of data between disk and computing system, and therefore the "smooth" job execution needed to make effective use of the compute resources.



Data transfers in Run 1 (ATLAS)



Looking for Opportunistic Resources

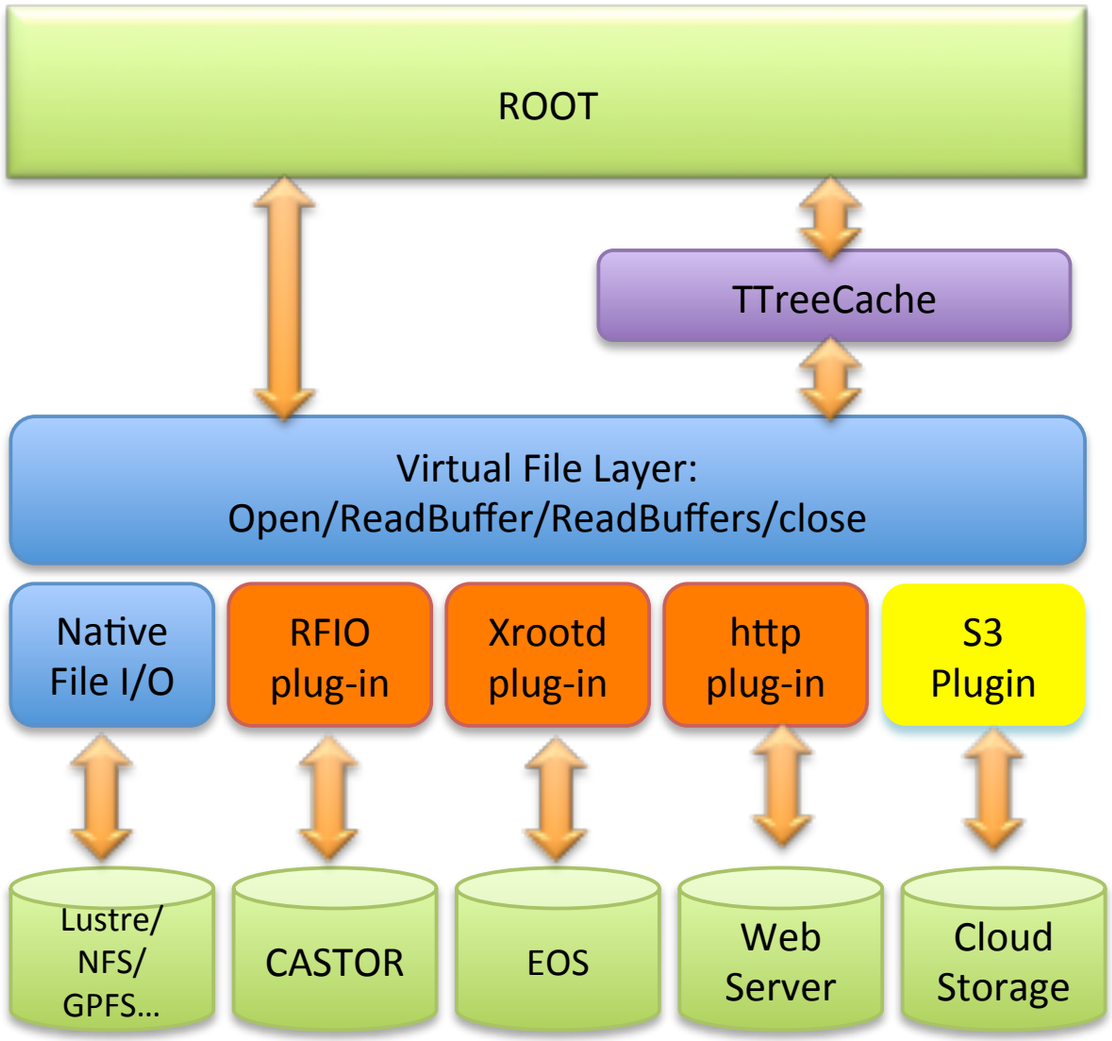
Slide: Borut Kersevan

- All our resource planning is done using **average** CPU (and disk) consumption rates.
  - And we are **using all available resources all the time** for diverse experiment activities.
  - The experiment analysis activities **peak** before big conference periods, can lead to congestions and backlogs in all the Grid demands.
- **Several venues to explore during Long Shutdown 1 (2013-2014):**
  - **Optimizing/changing our workflows, both in analysis and on the grid.**
    - It will necessarily involve also a change in the ways people analyze the data!
  - **Finding opportunistic resources:**
    - High Performance Computing centres have a lot of CPU available, we could use the available idle cycles for (a subset of) our activities,
      - e.g. MC event generation, possibly simulation.
    - Cloud resources: Again, for a subset of our activities, similar to HPC
      - If we are really hard pressed, even use commercial resources (?)
        - Exploring setups with Amazon EC2, Google Cloud ...
    - Opportunistic offers of big computing centers:
      - The experiments need to be able to simply and quickly integrate such resources into their distributed computing environment.
    - Volunteer computing resources: exploiting virtualization (CernVM), BOINC..
  - **Looking for solutions to speed up our code and accommodate our needs.**
- **A lot of activity foreseen in the experiment Software & Computing during LS1 to tackle this.**

Connecting Applications to Data - Access Protocols

CERN IT Department

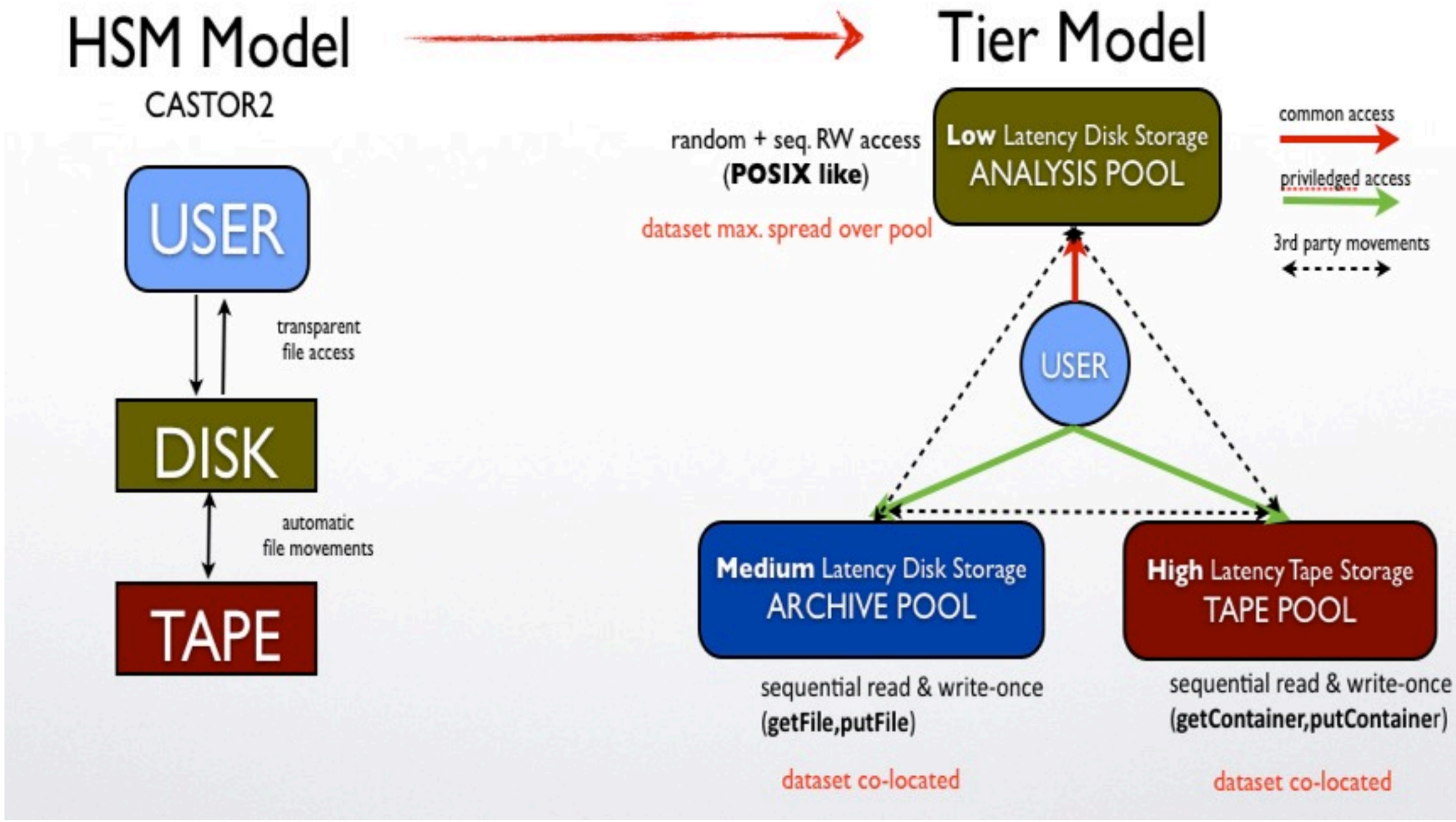
- LHC experiments use ROOT as persistency and analysis framework (-> Fons)
  - data access in a protocol agnostic way (plugins)
  - client side (in process) data cache
  - supports vector read and async read-ahead
- SRM as abstraction for storage administration
  - storage accounting and resource reservation
  - functionality progressively being integrated into access protocols



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Model Change: from HSM to User (=Experiment) controlled Data Tiers

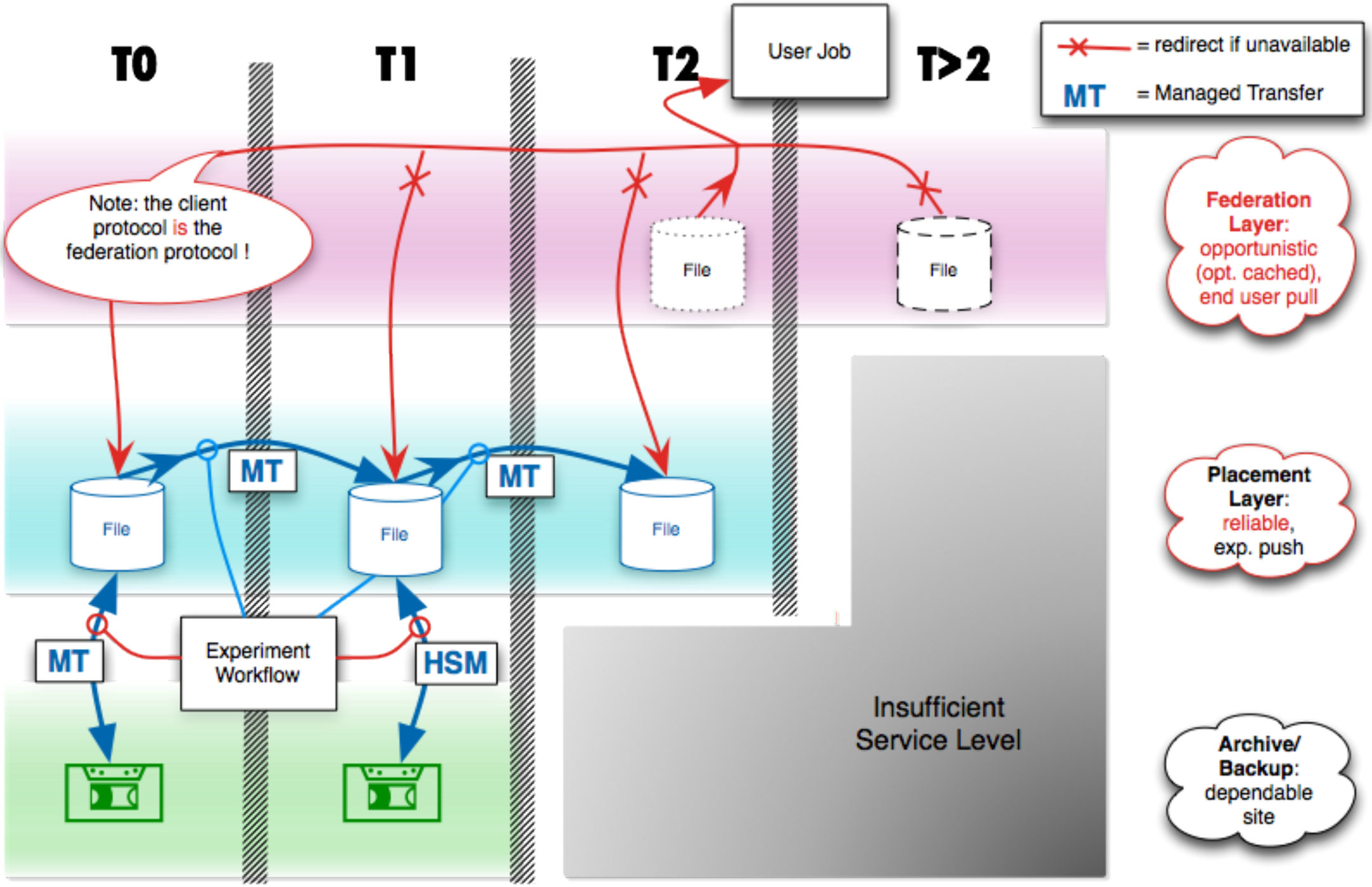
CERN IT Department



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

DSS Data Placement and Federation

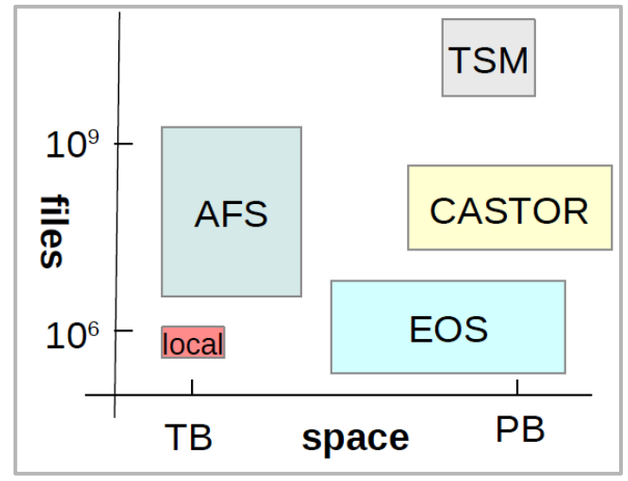
CERN IT Department



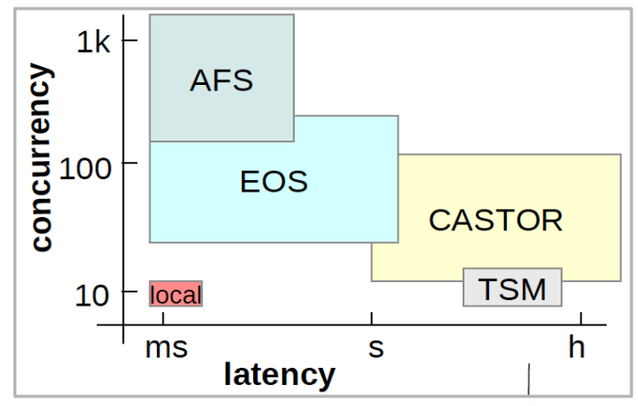
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Example Components: Physics Data Storage @ CERN

- CERN Data Centre just passed 100 PB
  - four LHC experiments
    - produced 75 PB in 3 years
    - raw measurements plus derived data
    - connected to 11 Tier 1 centres worldwide



  - Data Archive (CASTOR)
    - 88 PB from LHC and other experiments
    - focus: reliable, low $/TB, organised access



  - Analysis Disk Pools (EOS)
    - 16 PB of physics user or group data
      - 17'000 disks on 800 disk servers
    - focus: high speed, random access for many concurrent users

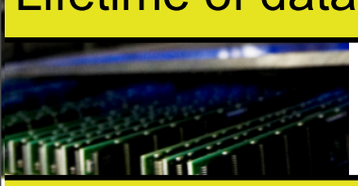
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CASTOR archive in Numbers

CERN IT Department

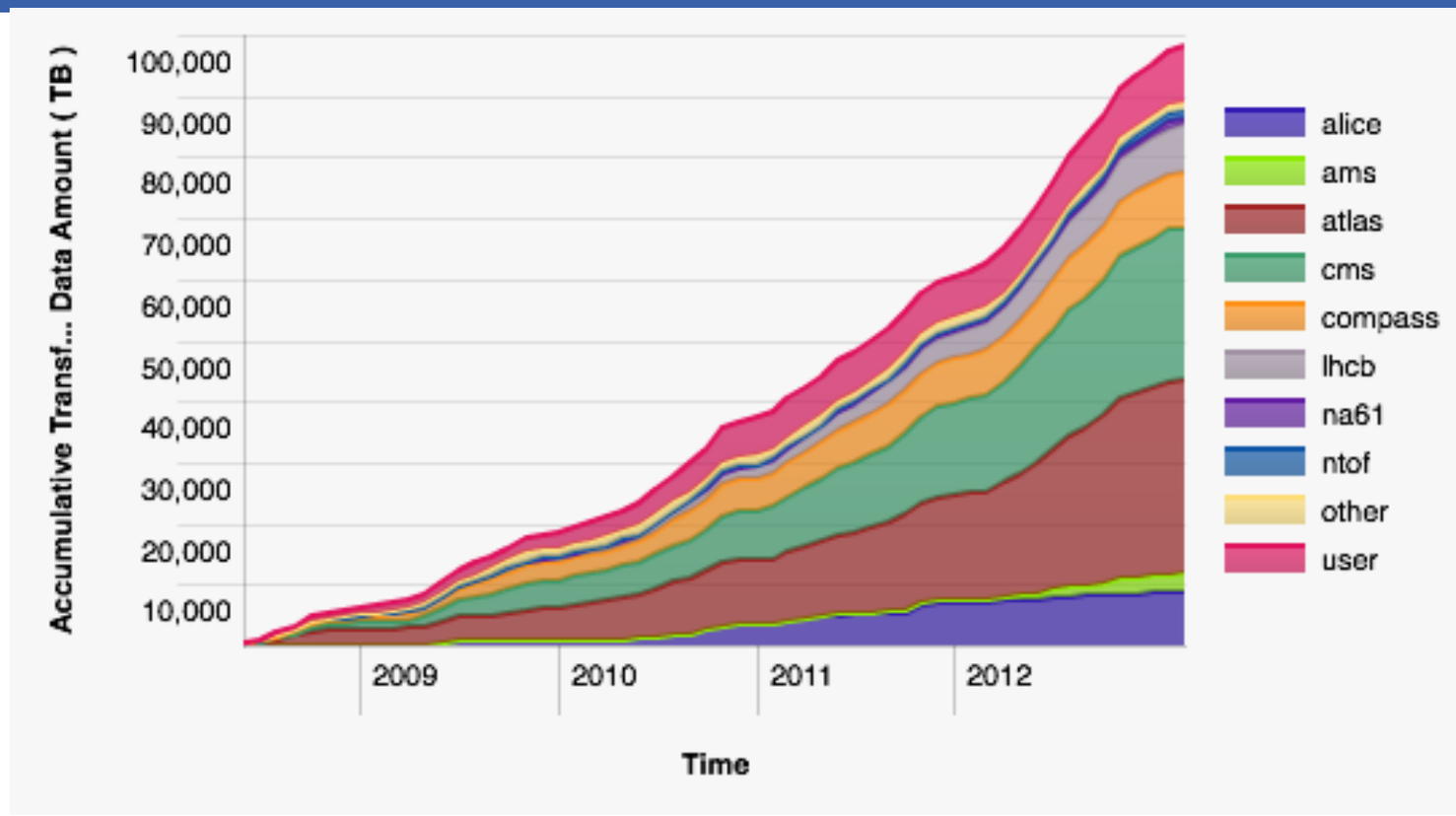
Data:
88PB (74PiB) of data on tape; 245M files over 48K tapes
Average file size ~360MB
1.5 .. 4.6 PB new data per month
Up to 6.9GB/s to tape during HI period

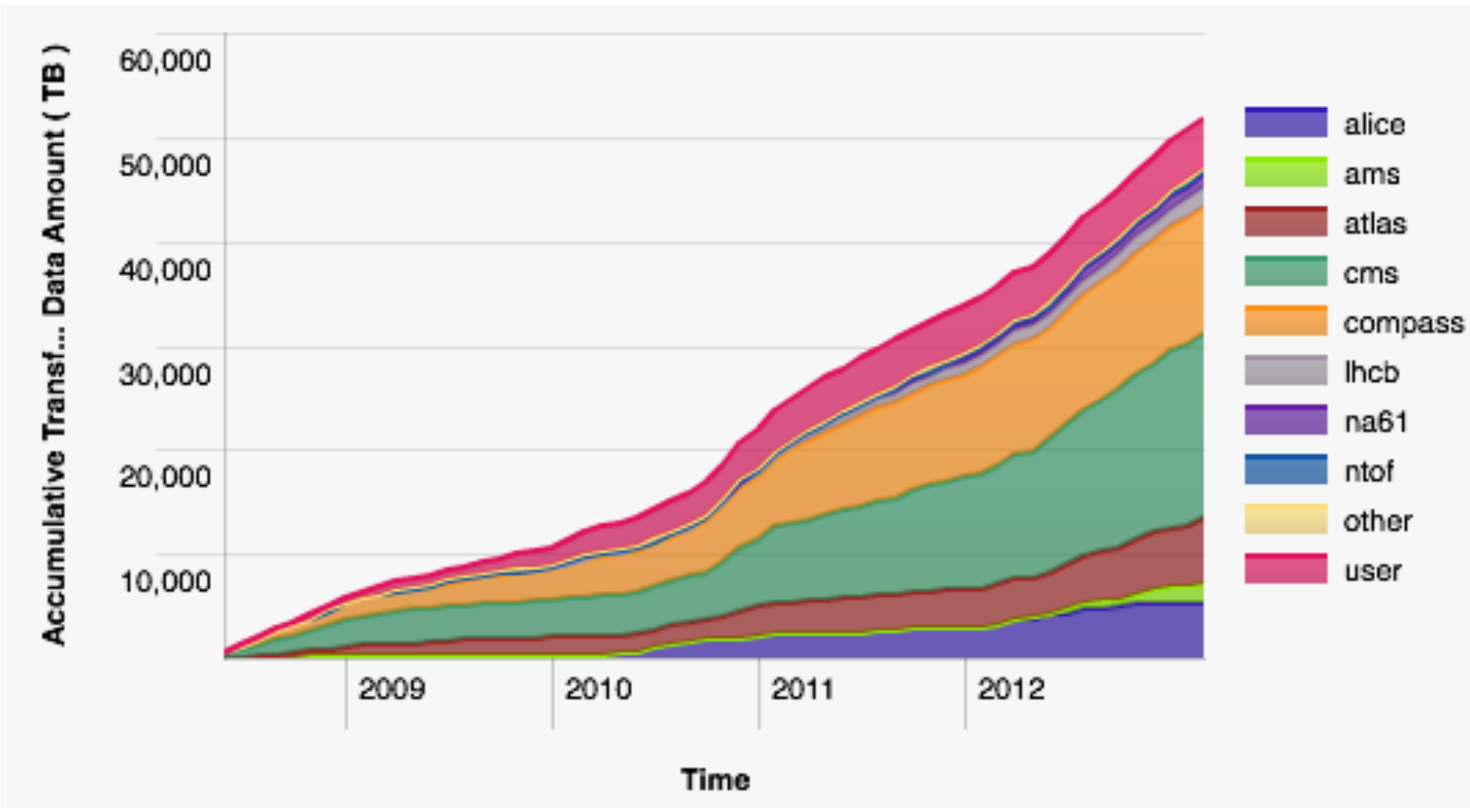
Lifetime of data: infinite



Infrastructure:
~ 52K tapes (1TB, 4TB, 5TB)
7 libraries (IBM and Oracle) – 65K slots
90 production + 20 legacy enterprise drives
15PB disk cache (staging + user access) on ~750 disk servers

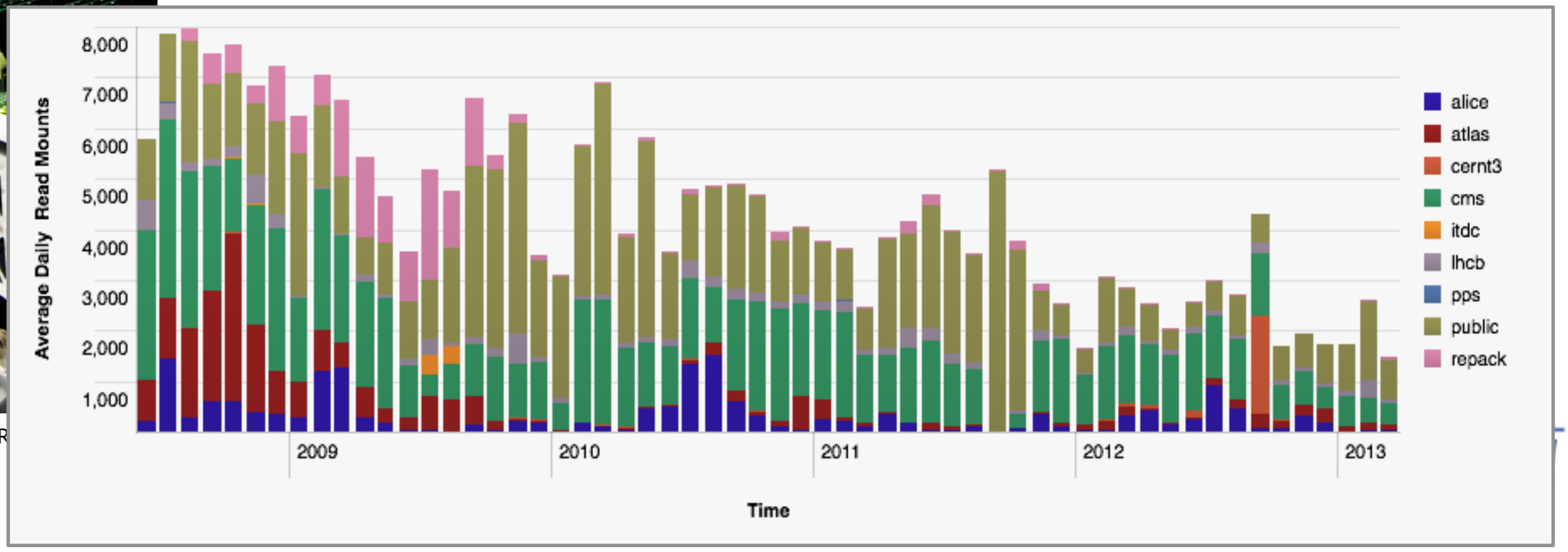
CASTOR
CERN Advanced STORage manager





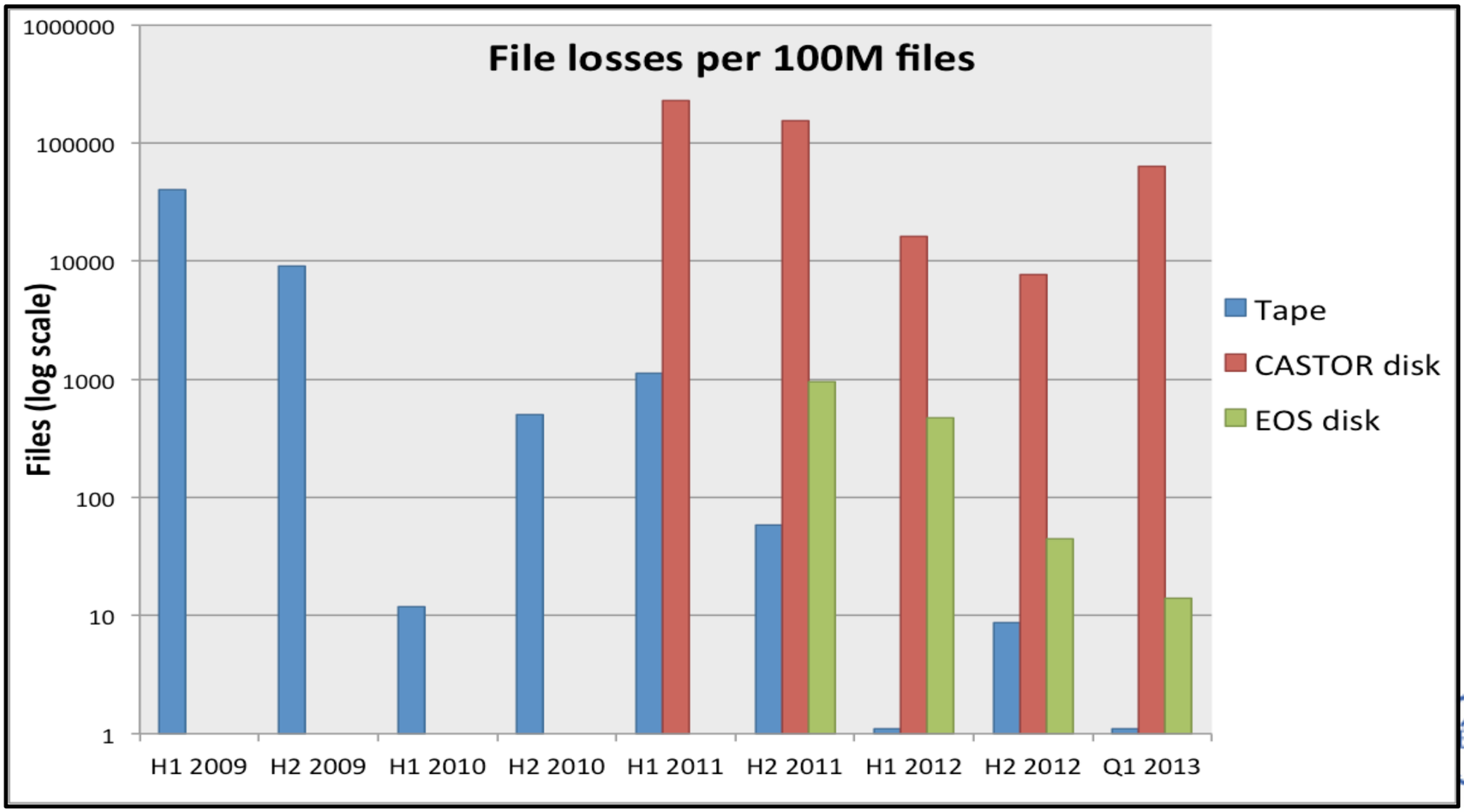
Tape mount rate reduction

- Deployed “traffic lights” to throttle and prioritise tape mounts
  - Thresholds for minimum volume, max wait time, concurrent drive usage, group related requests
- Developed monitoring for identifying inefficient tape users, encourage them to use bulk pre-staging on disk
- Work with experiments to migrate end-user analysis to EOS as mostly consisting in random access patterns
- Tape mount rates have decreased by over 50% since 2010, despite increased volume and traffic



Reliability

- File loss is unavoidable and needs to be factored in at all stages
- Good news: it has been getting better for both disk and tape
- Disk storage reliability greatly increased by EOS over CASTOR disk
  - RAID-1 does not protect against controller or machine problems, file system corruptions and finger trouble
- Tape reliability still ~O(1) higher than EOS disk
  - Note: single tape copy vs. 2 copies on disk



EOS Design Targets

CERN IT Department

- Project start: April 2010
- Focus: user analysis at CERN
  - many individual users with “chaotic” work patterns
  - many small output files, larger shared read-only input files
    - often only partial file access
    - many file seeks over “uninteresting” input events or branches
- Using xroot as client server framework
  - with an in-memory name space (no DB)
  - availability via file-level replication (configurable)
    - reduce operational effort at large volume scale

Pessimistic calculation assuming 1 MB file size

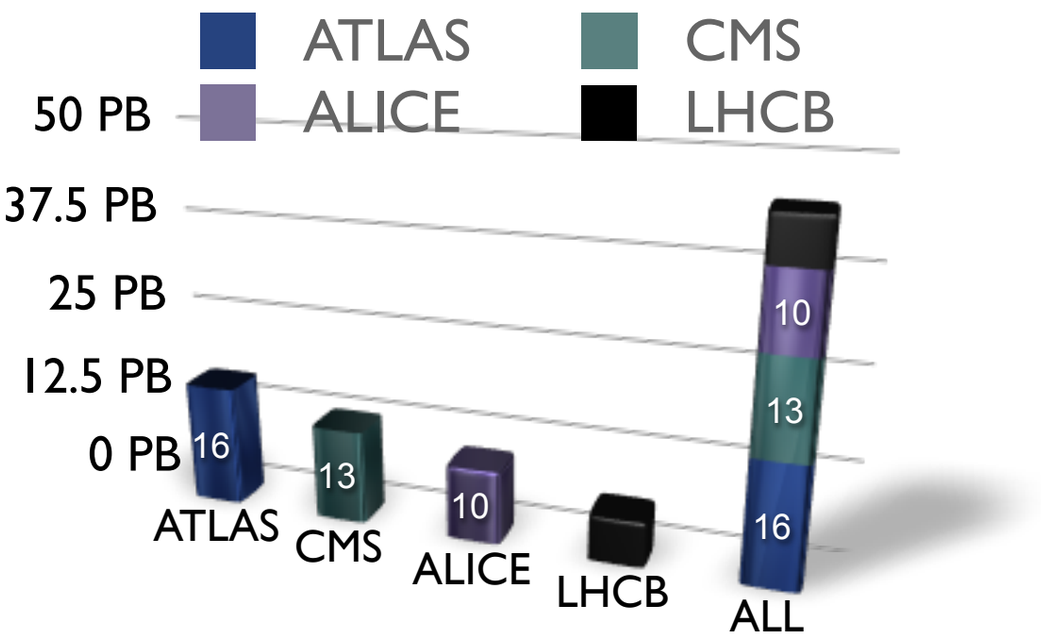
	Access Latency [s]	Files	File Container	Volume [bytes]
Analysis Pool	$10^{-3} – 10^{-2}$	$10^{9} – 10^{10}$ Billions	10^{6} - 10^{7} Millions	$10^{15} – 10^{16}$ Petabytes
Archive Pool	$10^{-2} – 10^{1}$	$10^{11} – 10^{12}$	10^{8} -10^{9} 100 Million+	$10^{17} – 10^{18}$ Exabytes
Tape Pool	$10^{1} – 10^{3}$	10^{11} - 10^{12}	10^{8} -10^{9} 100 Million+	$10^{17} – 10^{18}$ Exabytes

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

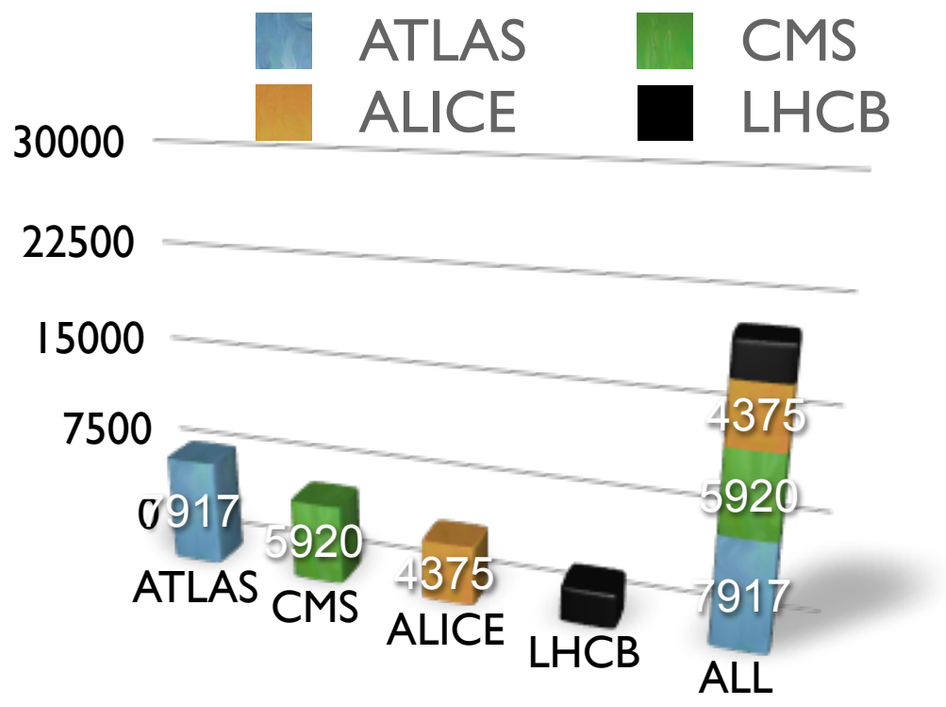
EOS Usage at CERN Today

CERN IT Department

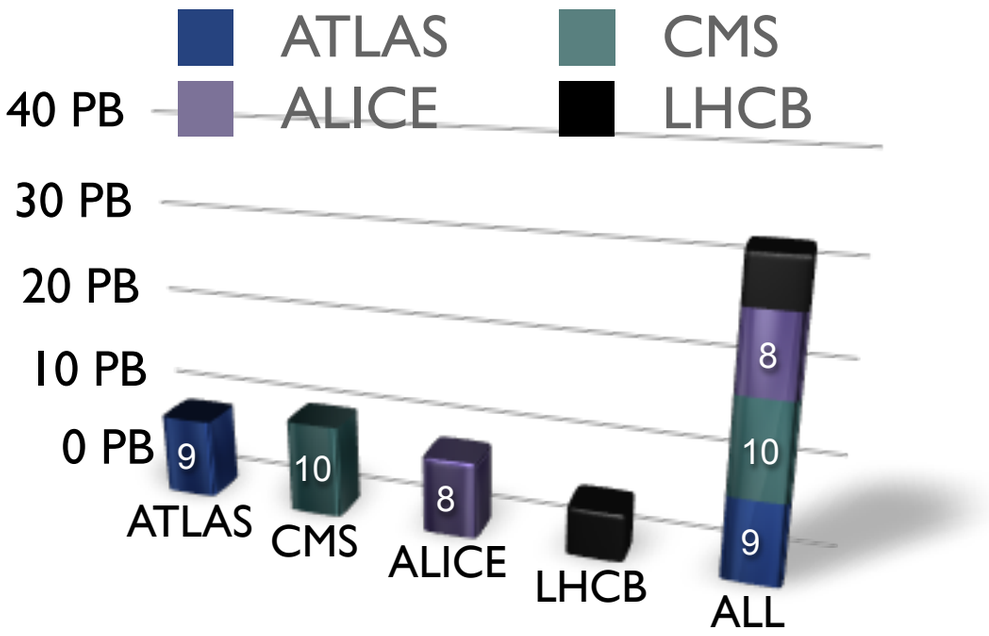
Raw Space 44.8 PB



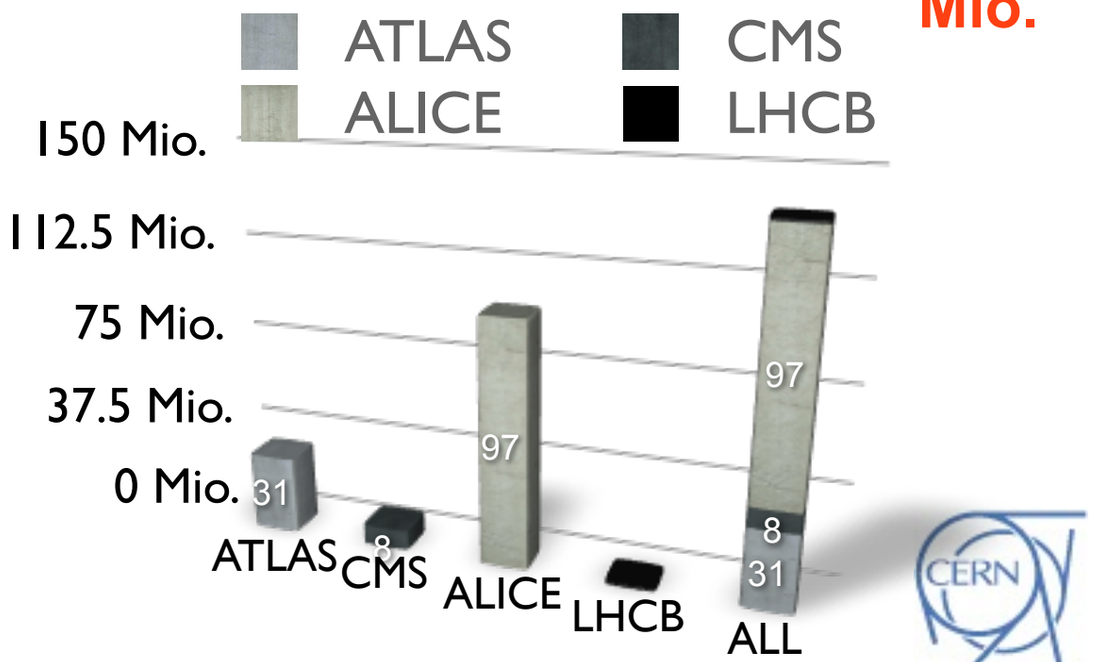
Harddisks 20.7k



Used Space 32.1 PB



Stored Files (Replicas) 136 (279) Mio.



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

New Technologies: Cloud Storage

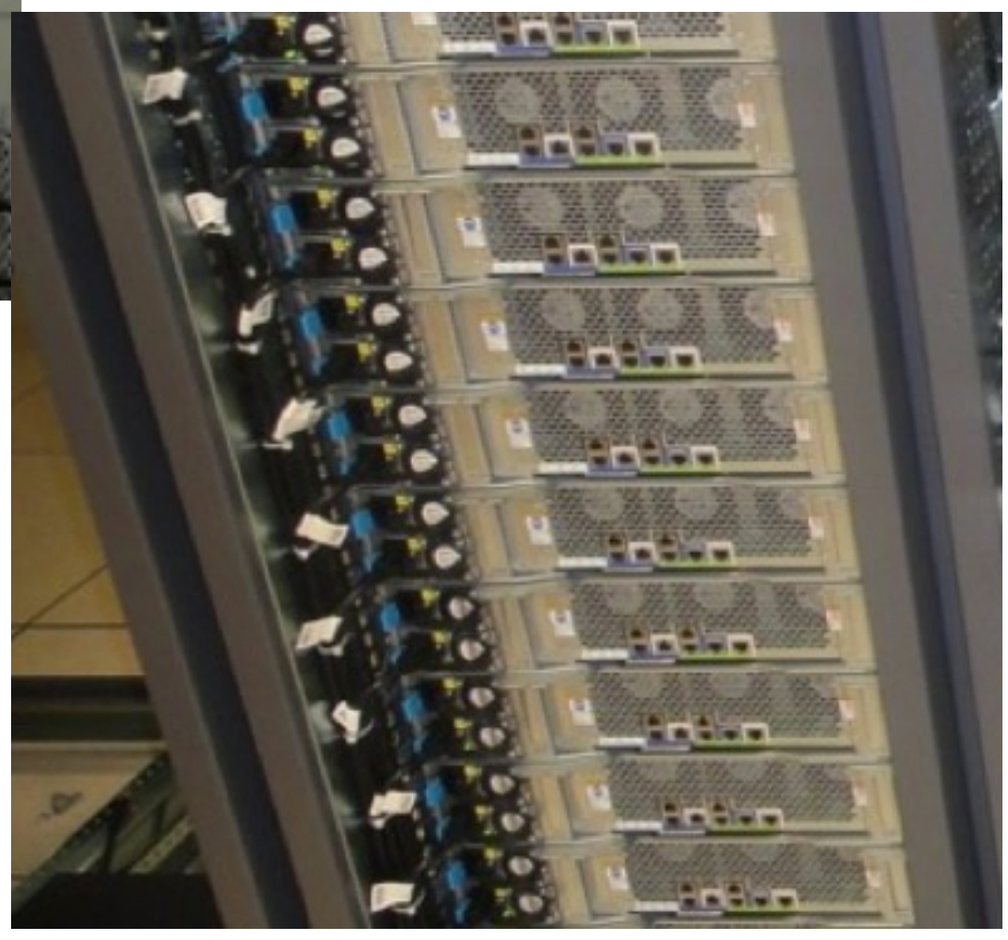
- Cloud computing and storage gain rapidly in popularity
  - Both as private infrastructure and as commercial service
  - Several investigations are taking place in the HEP and broader science community
- Evaluation goals
  - Changes in semantics, protocols, deployment model promise increased scalability at reduced TCO
  - Market is growing rapidly - need to understand if advantages can be confirmed with HEP work loads
  - Need to understand how cloud storage will integrate with (or change) current HEP computing models

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Example: Cloud Storage Appliances

CERN IT Department

- Huawei S3 storage appliance (0.8 PB)
  - logical replication
  - fail-in-place



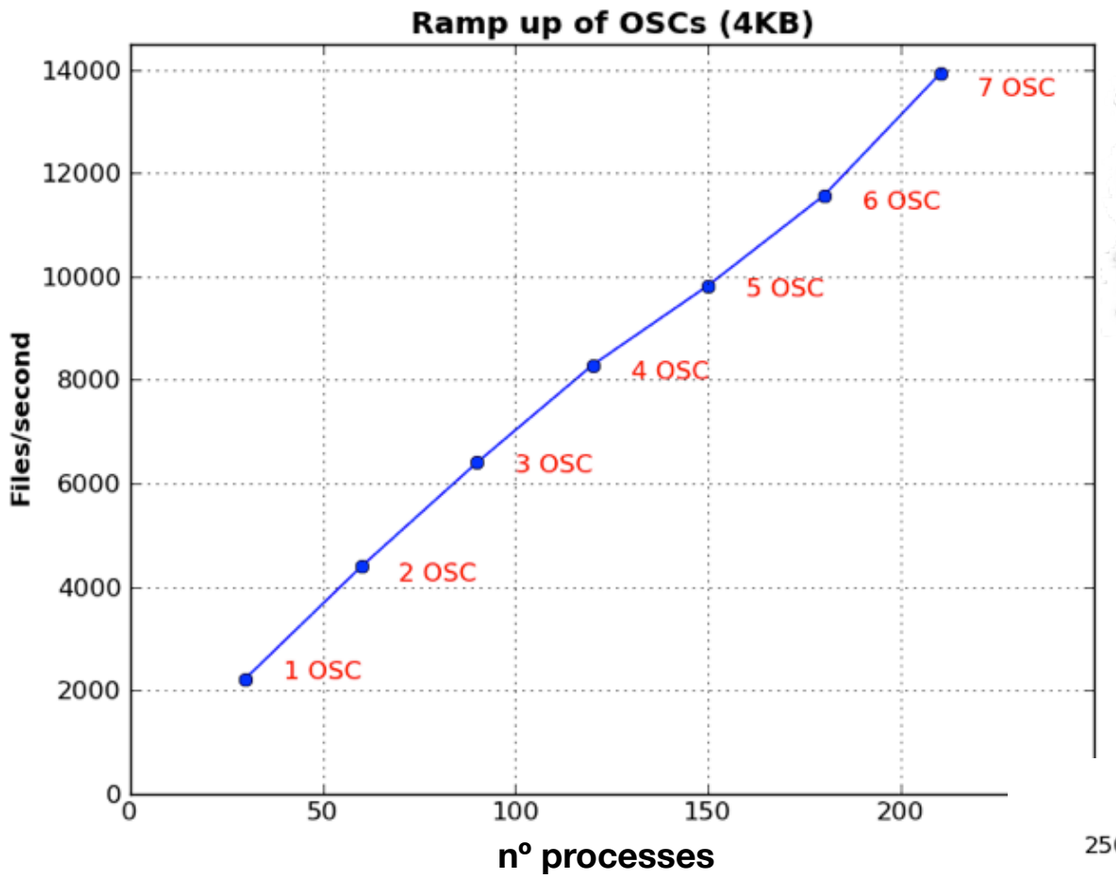
- CERN openlab
  - joint project since Jan ’12
  - Testing scaling and TCO gains with prototype applications

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Meta-data & Bandwidth Scalability

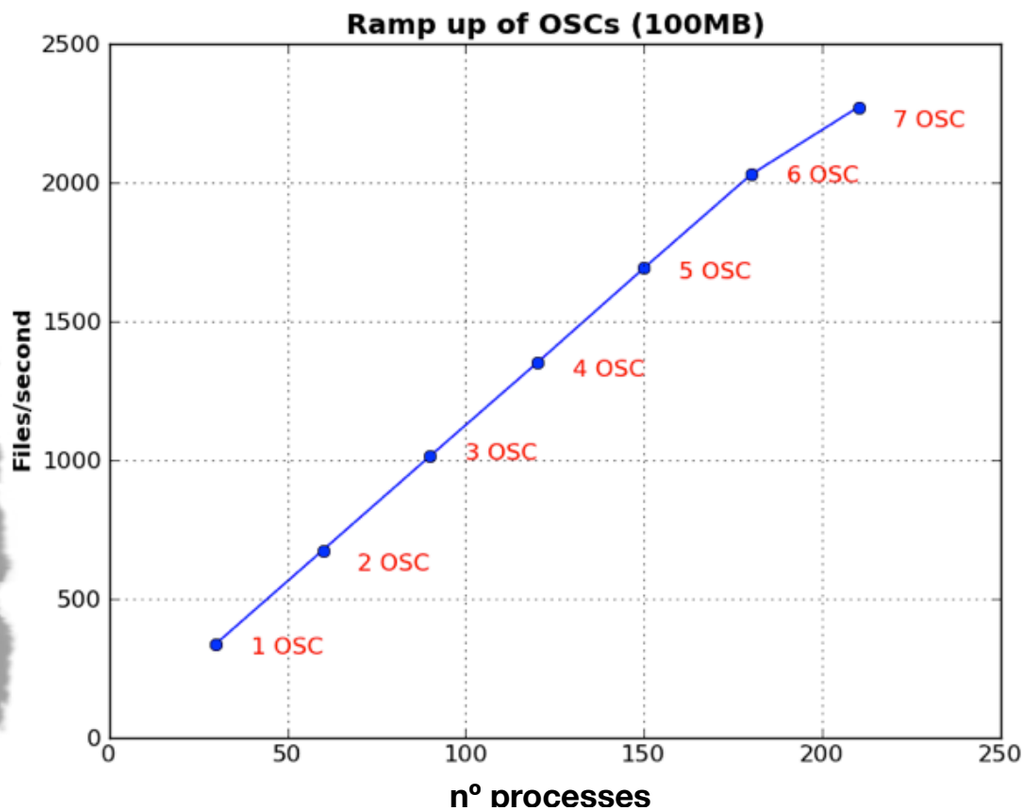
CERN IT Department



Each box processes around 2000 files/sec

Each box handle around 350MB/second

-> linear scaling !



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Trends & Opportunities

- Move from disk-based DBs to real-memory will be possible for an increasing number of services / appl's
  - new random access memory technologies will accelerate this trend
  - significant change for
    - commercial products and markets
    - application access methods
  - independence from access protocol will stay important to take advantage
- Traditional "local disk <-> local archive" coupling is being challenged by inter-site movements
  - number of archives sites will be further consolidated
    - tape is unlikely to disappear soon for reasons of economy and trust
  - archive use case (including media migration) looks similar between different science communities

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Trends & Opportunities

CERN IT Department

- Cloud & BigData technologies will continue to raise
  - even if often for non-technical reasons
    - eg interest in Hadoop still exceeds the available evidence of gains
  - students are more likely to have hadoop experience then home-grown technologies
  - exploiting short term resource offers will require compatibility with commercial interfaces (cloud bursting)

- The science share of the spent computing budget will continue to decrease. Market forces will
  - push science to use fewer, but more sustainable products
  - increase commonalities between different scientific computing areas

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

DSS

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it